The Knox Method and Other Tests for Space-Time Interaction

Martin Kulldorff (1),* Ulf Hjalmars (2)
(1) Division of Biostatistics, Department of Community Medicine and Health Care, University of Connecticut School of Medicine, Farmington, CT; (2) Department of Pediatrics, Östersund Hospital, Östersund, Sweden

Abstract

The Knox method, like other tests for space-time interaction, is biased in situations in which there are geographical population shifts; that is, when there are different percentages of population growth in different regions. In this paper, the size of the population shift bias is investigated for the Knox test, and it is shown that it can be a considerable problem. This paper then presents a Monte Carlo method for constructing unbiased space-time interaction tests, illustrating the method for the Knox test and for a combined Knox test. Practical implications are discussed in terms of the interpretation of past results and the design of future studies.

Keywords: bias, Jacquez' test, Knox test, Mantel's test, population shifts

Introduction

Space-time interaction tests are used to evaluate whether there is space-time clustering of events after purely spatial and purely temporal clustering are adjusted for. These tests are frequently applied in epidemiological studies, in which it is of interest to know whether cases of some disease are more clustered than would be expected based on the underlying geographical population distribution and on any purely temporal trend. Two excellent surveys on space-time interaction tests have been written by Mantel (1) and Williams (2). Comparative evaluations and power studies have been done by Chen, Mantel, and Klingberg (3) and by Jacquez (4).

The most widely used statistical technique for testing space-time interaction was developed by Knox (5). In the Knox test, the time and geographical location of each case are noted, and the distance between each possible pair of cases is calculated in terms of both time and space. If many of the cases that are "close" in time are also "close" in space ("close" is defined by the user), or vice versa, then there is space-time interaction. This could be an indication that a disease is infectious or that it is caused by some other type of agent that appears locally at specific times, such as food poisoning.

In a survey of epidemiological articles published between 1960 and 1990, Daniel Wartenberg and Michael Greenberg (6) found 59 different studies that used the Knox method. Many of these were concerned with leukemia, and the results from such studies have been used as evidence supporting a viral etiology of the disease (7,8).

The Knox test is an elegant and, in many ways, attractive method. For example, it

* Martin Kulldorff, University of Connecticut School of Medicine, 263 Farmington Avenue, Farmington, CT 06030-6325 USA; (p) 860-679-5473; (f) 860-679-5464; E-mail: martink@cortex.uchc.edu. Note: this work was conducted while this author was at the Biometry Branch, Division of Cancer Protection, of the National Cancer Institute, Bethesda, MD.

is simple and straightforward to calculate the test statistic, and using the test requires knowledge only of cases, not controls. There is, however, a well-known problem with the method.

Mantel (1) pointed out that the Knox test is biased if the rate of population growth is not constant for all geographic sub-areas. We call this the *population shift bias.* Shifts in the population distribution create space-time interaction among any random sample of individuals, including sets of cases generated under the null hypothesis of equal disease risk. The Knox statistic is constructed so as to pick up any type of space-time interaction; it does not distinguish whether that interaction is due to shifting population distributions or to some disease-related phenomenon. This is not a flaw of the test per se, and is not a problem if one is looking for any type of space-time interaction. However, interest is typically focused—as in epidemiology—on disease-related phenomena, not shifts in population distribution, so the latter should be adjusted for.

While the existence of the population shift bias has long been known, the magnitude of the bias has not been studied for any real datasets, and the bias has typically been ignored in practical applications. In the "Estimation of the Population Shift Bias" section, the bias of the ordinary Knox test is estimated for two different datasets: the child population in Sweden from 1976 to 1994 (a fairly stable population) and the total population in New Mexico (where there have been large population shifts) from 1973 to 1991. The estimations show that the bias is considerable for some cases.

Klauber and Mustacchi (9) suggested that the population shift bias could be reduced by dividing the data into several parts corresponding to different time periods. Within these parts, the population would be more stable. A test statistic would then be calculated separately for each part, and the statistics would be summed to get an overall test. This method reduces the bias but does not eliminate it. Unfortunately, it also decreases the power of the test; pairs of cases falling in different data parts would not be used, leading to loss of information.

A simple unbiased version of the Knox test is presented in the section entitled "An Unbiased Knox Test." This test adjusts not only for purely spatial and purely temporal variations, but also for the space-time interaction inherent in the background population. It does so without the loss of power associated with the Klauber-Mustacchi approach. Its one drawback is that it requires knowledge of the underlying population distribution.

While this paper is focused on the Knox method, which is the most commonly used space-time interaction test, other space-time interaction tests suffer from the same population shift bias. This includes the methods proposed by David and Barton (10), Mantel (1), Pike and Smith (11), Diggle et al. (12), Jacquez (4), and Baker (13). This paper's approach for constructing an unbiased Knox test can also be used to construct unbiased versions of these other methods.

A second issue with the Knox method relates to the choice of critical distances to define which pairs of cases are close in space and time respectively. Unless the investigator has a fairly clear idea of the scale at which clustering may occur, this is a problem. Separate tests are often performed for a number of different critical distances (e.g., Gilman and Knox, 1995 [14]). It is possible to do a Bonferroni-type adjustment for the multiple testing inherent in such a procedure, but because the test statistics calculated for adjacent critical distances are highly correlated, there is loss of power when using such a method. In practice it is seldom used. Baker (13) has presented a combined Knox

test, providing a single hypothesis test with multiple critical distances. The approach presented in the section entitled "An Unbiased Combined Knox Test" uses the same basic idea to deal with multiple testing.

If the simple modification to the Knox test described here were implemented in actual studies, the value of those studies would greatly increase. There would no longer be any uncertainty about whether a significant result is due simply to shifts in the geographical population distribution, and there would be no issue of multiple testing. The Knox test is an intuitive, elegant method. With its major weaknesses resolved, we hope, it will continue to be used for years to come.

The Knox Test

Let n be the total number of cases, so that there are $N=n(n-1)/2$ distinct pairs of cases. Let N_t be the number of case pairs that are closer to each other in time, compared to some specified temporal distance. Likewise, let N_s be the number of pairs close in space as defined by some geographic distance. Finally, let X be the number of case pairs that are close both in time and space.

The observed value of X is the test statistic of the Knox method (5). To adjust for purely spatial and purely temporal inhomogeneities in the data, the test statistic is evaluated conditionally on N_t and N_s. Under the null hypothesis of no space-time interaction, the expected value of X is $E[X \mid N_t,N_s]=N_tN_s/N$ (15).

Knox (5) conjectured that X is approximately Poisson-distributed. Barton and David (15) showed this to be true when N_t and N_s are small compared to N, in the sense that the variance of X is then approximately equal to its expected value. More importantly, by application of graph theory and by also conditioning on the second-order terms, they obtained an exact formula for the variance:

$$V[X|N_t, N_s, N_{2s}, N_{2t}] =$$

$$\frac{N_sN_t}{N} + \frac{4N_{2s}N_{2t}}{n(n-1)(n-2)} + \frac{4[N_s(N_s-1)-N_{2s}][N_t(N_t-1)-N_{2t}]}{n(n-1)(n-2)(n-3)} - \left(\frac{N_sN_t}{N}\right)^2$$

where N_{2s} is the number of pairs of case pairs close in space that have one case in common, and where N_{2t} is defined equivalently for time.

In practical applications, different approximations of the test statistic's distribution have been used. The Cluster software package, written by Aldrich and Drane (16), uses the Poisson approximation, as originally proposed by Knox (5). Gilman and Knox (14) and many others have done likewise, except that they have used the normal approximation for the Poisson distribution, keeping the variance equal to the mean. We will call this approach the *Poisson-based approximation*. An alternative approach is to use a normal approximation with the mean and variance given by Barton and David (15). We will call this the *Barton-David-based approximation*. Yet another option, originally proposed by Mantel (1), is to use Monte Carlo hypothesis testing (17) by permuting the times among the fixed spatial locations. This is implemented as part of the Stat! software package (18); Petridou et al. (8) provide one example of its use.

Before estimating the population shift bias, as in the next section, it is important to look at any potential bias due to the distributional assumptions of the Knox test

statistic. Table 1 contains bias estimates for the Poisson- and Barton-David-based approximations when the Knox test is applied to a hypothetical child population in Sweden. For all years from 1976 to 1994, the population is artificially fixed at the 1982 level so that there are no population shifts. The data are aggregated into 2,507 parishes. The parish and month were randomly selected for each of 1,000 and 10,000 cases in proportion to the 1982 population for each parish, and in proportion to the length of each month.

When N_t and N_s are small compared to N, the Poisson-based approximation works well. When N_t and N_s are larger, though, there is some bias. This is as expected based on the theoretical results of Barton and David (15). The Barton-David-based approximation, on the other hand, works well across the board for the Swedish data. This is important to remember when estimating the population shift bias, as in the next section. By definition, the Monte Carlo procedure provides an unbiased test when there are no shifts in the population distribution.

Estimation of the Population Shift Bias

Differential population growth can be caused by internal migration between different regions, by geographically differential emigration or immigration rates, or by geographically differential birth or death rates. If the disease risk is related to age, the bias can also be caused by different age structures in different regions, whether that structure changes over time or not; as the population ages, the age-specific population counts change over time to different degrees in different regions.

The magnitude of the population shift bias of any test for space-time interaction depends on the specific geographic area and time period under study. In general, shorter overall time periods result in less bias because there is less time for population shifts to occur, as pointed out by Klauber and Mustacchi (9). Nothing general can be said about specific geographic areas. To give some idea of the extent of the bias, we have calculated the population shift bias of the ordinary Knox test for two different datasets.

The first dataset is the child population in Sweden from 1976 to 1994, aggregated to the 2,507 parishes. The second dataset is the total New Mexico population from 1973 to 1991, aggregated to 32 counties. (The second dataset is available at http://dcp.nci.nih.gov/BB/datasets.html.) For the New Mexico dataset, Cibola and Valencia are counted as one county for the whole time period even though they became two different counties in 1981. The geographic distance between cases is the distance between the parish/county centroids to which they belong. When the critical geographic distance is 0, only those cases located in the same parish are considered spatial neighbors. For both datasets, the case times are noted in months. When the critical temporal distance is zero months, neighboring cases are only those occurring in the same calendar month; when it is three months, neighboring cases are those occurring in months at most three calendar months apart (e.g., January and April, but not January and May); and so on.

To put these datasets in a proper context, the population growth for various subregions is provided in Table 2. For the child population in Sweden, Table 2 shows the population growth in each of the country's 24 counties, or läns. Table 2 shows only part of the picture, though; the data were analyzed at the much finer level of 2,507 parishes. The percentage of change, naturally, varies more for the smaller parishes. The 470

Table 1 Estimated True Significance Levels for the Ordinary Knox Test When Applied to the Childhood Population in Sweden

		Poisson Approximation				Barton-David Approximation			
		$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.01$	
cut-off points		# cases		# cases		# cases		# cases	
km	months	1000	10000	1000	10000	1000	10000	1000	10000
0	0	.064	.072	.021	.018	.064	.073	.021	.019
0	3	.052	.052	.013	.010	.054	.056	.014	.011
0	6	.051	.050	.010	.006	.055	.056	.012	.009
0	12	.044	.046	.009	.008	.052	.053	.012	.010
0	24	.034	.044	.006	.007	.050	.051	.011	.009
2	0	.064	.075	.020	.018	.064	.076	.020	.018
2	3	.053	.053	.013	.010	.055	.054	.014	.011
2	6	.048	.050	.011	.007	.053	.055	.013	.008
2	12	.043	.044	.009	.008	.052	.050	.012	.010
2	24	.035	.041	.006	.007	.050	.050	.012	.010
5	0	.062	.059	.020	.012	.063	.060	.020	.012
5	3	.052	.056	.012	.009	.054	.058	.014	.010
5	6	.047	.051	.011	.010	.052	.055	.013	.012
5	12	.047	.052	.009	.011	.055	.053	.013	.011
5	24	.040	.062	.007	.014	.054	.050	.013	.011
10	0	.055	.054	.015	.013	.056	.056	.015	.013
10	3	.049	.058	.012	.012	.052	.058	.012	.012
10	6	.047	.056	.010	.010	.053	.053	.011	.010
10	12	.049	.067	.010	.018	.056	.051	.013	.011
10	24	.046	.118	.011	.050	.053	.050	.014	.012
20	0	.052	.054	.013	.010	.055	.056	.014	.011
20	3	.049	.057	.011	.011	.053	.056	.013	.010
20	6	.049	.069	.010	.017	.055	.056	.012	.010
20	12	.054	.109	.015	.039	.059	.056	.016	.015
20	24	.064	.196	.019	.120	.058	.055	.017	.012
50	0	.049	.051	.010	.010	.053	.056	.012	.012
50	3	.047	.056	.011	.011	.054	.057	.014	.012
50	6	.046	.070	.011	.019	.052	.055	.014	.013
50	12	.049	.116	.013	.049	.053	.054	.015	.012
50	24	.067	.221	.022	.139	.058	.051	.017	.011

Note: Cases were randomly generated according to the 1982 population, so there is no population shift bias. The bias due to the Poisson and Barton-David approximations for the distribution of the test statistics is the difference between the numbers reported and the nominal significance level.

Table 2 Population Changes in the 24 Läns of Sweden and in the 32 Counties of New Mexico

Sweden				New Mexico			
	child population				total population		
län	1976	1994	change	county	1973	1991	change
Stockholm	319901	335044	+5%	Bernalillo	353813	490248	+39%
Uppsala	54445	60829	+12%	Catron	2372	2507	+6%
Södermanland	57765	52666	-9%	Chaves	45204	58699	+30%
Östergötland	86058	83198	-3%	Colfax	12577	12743	+1%
Jönköping	68603	64988	-5%	Curry	42709	44613	+4%
Kronoberg	38228	36530	-4%	DeBaca	2509	2310	-8%
Kalmar	51718	48694	-6%	Doña Ana	76915	140696	+83%
Gotland	12389	12426	0%	Eddy	40940	49998	+22%
Blekinge	34501	29280	-15%	Grant	23549	27986	+19%
Kristianstad	60617	59318	-2%	Guadalupe	4889	4102	-16%
Malmöhus	157924	155186	-2%	Harding	1234	987	-20%
Halland	52084	56385	+8%	Hidalgo	5108	5937	+16%
Göteborg-Bohus	149254	146911	-2%	Lea	48907	55584	+14%
Älvsborg	96494	94708	-2%	Lincoln	8395	12824	+53%
Skaraborg	60117	59042	-2%	Los Alamos	15315	17908	+17%
Värmland	58583	55198	-6%	Luna	13493	18984	+41%
Örebro	58775	54432	-7%	McKinley	46826	62746	+34%
Västmanland	61058	52042	-15%	Mora	4712	4208	-11%
Kopparberg	59303	58938	-1%	Otero	42303	52256	+24%
Gävleborg	61696	55583	-10%	Quay	10980	10564	-4%
Västernorrland	56258	49316	-12%	Rio Arriba	27339	34330	+26%
Jämtland	27216	26881	-1%	Roosevelt	16477	17258	+5%
Västerbotten	52376	54642	+4%	Sandoval	23858	65975	+177%
Norrbotten	63038	53520	-15%	San Juan	58718	94028	+60%
				San Miguel	23452	26074	+11%
				Santa Fe	61250	101675	+66%
				Sierra	7976	10098	+27%
				Socorro	10492	14696	+40%
				Taos	18053	23679	+31%
				Torrance	5730	10658	+86%
				Union	5060	4136	-18%
				Valencia	43192	70135	+62%
Total	1798401	1755757	-2%	Total	1104347	1548642	+40%

Note: For the Swedish data, the actual analysis was done using much less aggregated data.

parishes with more than 1,000 children in 1976 had an average population decrease of 2.5% from 1976 to 1994, with a standard deviation of 30.4 percentage points. The equivalent standard deviations for other subgroups were 33.3 for 275 parishes with 1976 populations in the 500–1000 range, 31.4 for 486 parishes in the 200–500 range, 72.0 for 467 parishes in the 100–200 range, and 122.8 for 809 parishes with 1976 populations of less than 100. The population growth in New Mexico is also presented in Table 2. Between

1973 and 1991, one county's population doubled while many other counties had a fairly constant population.

To estimate the population shift bias, cases were randomly assigned to a parish (or county, for New Mexico) and to a particular month with probability proportional to the actual population in that parish during that month. In this way, the cases were randomized with population shifts taken into account. The population for a particular month was obtained through linear interpolation, using yearly population data for New Mexico and the years 1976, 1982, 1988, and 1994 for Sweden. Separate calculations were done for 1,000, 4,000, and 10,000 randomized cases. For each random Monte Carlo replication of the fixed number of cases, the test statistic was calculated and compared with its nominal critical region using the Barton-David distributional approximation. Without bias, 5% of the test statistics from the Monte Carlo replications should fall within the critical region. The actual numbers are given in Tables 3 and 4.

The population shift bias for the Swedish data is the difference between the numbers reported in Table 3 and those reported in Table 1 for the Barton-David approximation. The total bias is the difference between Table 3 and the nominal significance levels. For the Swedish data, there is very little bias using the original Knox test when the total number of cases observed is 1,000. With more cases, the bias increases. It is a considerable problem with 10,000 cases observed.

For New Mexico, the bias is considerable for 1,000, 4,000, or 10,000 cases, as can be seen from Table 4. Note that it is not the total population increase of 40% that causes the bias. If the increase were the same in all counties, the population shift bias would be zero.

The bias estimates in Tables 1, 3, and 4 were calculated using 20,000 random replications of the fixed number of cases. The 95% confidence intervals are ±0.007 when the estimate is around 0.50, and ±0.003 when the estimate is around 0.05. If the Poisson approximation is used instead of the Barton-David approximation, the total bias is about the same or higher (not shown), as would be expected considering Table 1.

As Tables 3 and 4 show, the population shift bias increases with an increased number of total cases observed. Why? By definition, the population shift bias is the probability that a method will detect space-time interaction due to the population shift when there is no space-time interaction of any other kind. That is, a method's population shift bias is identical to its power to detect a population shift using a random sample from the population. The larger the random sample, the greater the power; by consequence, the more cases, the bigger the population shift bias. In a sense, this is a Catch-22 situation. We could reduce the population shift bias by analyzing a smaller number of cases, but that would also reduce the power to detect space-time interaction due to any biological phenomena of interest.

The population shift bias also varies with the choice of critical geographical distance. Such differences are data-dependent. Consider a situation in which the child population over time is identical in several cities, but in which, within those cities, there is a continuous child population shift. New suburbs have many small children who grow older together with the suburbs until they move out and leave a predominantly adult population behind. This will lead to a population shift bias for small values of the critical geographic distance, but not necessarily for large ones. On the other hand, increased critical distances will result in more space-time case pairs, increasing the power

Table 3 Estimated True Significance Levels for the Ordinary Knox Test Using the Barton-David Approximation, When the Nominal Levels Are α=0.05 and α=0.01, for the Childhood Population in Sweden, 1976–1994

		$\alpha = 0.05$			$\alpha = 0.01$		
cut-off points		number of cases			number of cases		
km	months	1000	4000	10000	1000	4000	10000
0	0	.069	.074	.105	.021	.019	.032
0	3	.062	.079	.141	.016	.020	.040
0	6	.063	.087	.181	.018	.021	.064
0	12	.062	.102	.246	.018	.026	.101
0	24	.066	.126	.299	.018	.033	.132
2	0	.066	.072	.105	.021	.018	.034
2	3	.063	.086	.155	.017	.022	.041
2	6	.063	.091	.198	.017	.026	.071
2	12	.064	.105	.267	.017	.029	.105
2	24	.068	.138	.331	.018	.039	.146
5	0	.065	.070	.102	.018	.018	.029
5	3	.059	.087	.164	.016	.022	.050
5	6	.059	.100	.214	.016	.030	.077
5	12	.063	.114	.240	.016	.036	.097
5	24	.065	.116	.231	.017	.036	.090
10	0	.056	.061	.088	.015	.015	.020
10	3	.057	.081	.134	.014	.021	.043
10	6	.058	.092	.166	.015	.025	.058
10	12	.060	.098	.158	.017	.031	.059
10	24	.061	.092	.108	.017	.032	.037
20	0	.054	.061	.081	.013	.016	.019
20	3	.058	.071	.100	.015	.020	.030
20	6	.058	.082	.108	.016	.024	.032
20	12	.060	.075	.087	.018	.023	.024
20	24	.059	.066	.054	.016	.020	.014
50	0	.051	.055	.059	.013	.013	.014
50	3	.058	.062	.080	.014	.016	.022
50	6	.056	.068	.082	.015	.023	.024
50	12	.061	.064	.063	.016	.019	.019
50	24	.057	.058	.050	.017	.013	.015

Note: The difference between these numbers and those in Table 1 is the population shift bias.

Table 4 Estimated True Significance Levels for the Ordinary Knox Test Using the Barton-David Approximation, When the Nominal Levels Are α=0.05 and α=0.01, for the Population in New Mexico, 1973–1991

		$\alpha = 0.05$			$\alpha = 0.01$		
cut-off points		number of cases			number of cases		
km	months	1000	4000	10000	1000	4000	10000
0	0	.064	.085	.101	.016	.023	.027
0	3	.071	.098	.148	.020	.028	.048
0	6	.076	.105	.155	.022	.034	.049
0	12	.077	.110	.151	.023	.036	.051
0	24	.078	.104	.134	.025	.033	.041
50	0	.069	.127	.240	.019	.039	.088
50	3	.091	.212	.450	.030	.081	.232
50	6	.109	.239	.469	.036	.100	.250
50	12	.114	.241	.444	.042	.103	.227
50	24	.115	.208	.353	.041	.084	.159
100	0	.072	.157	.370	.019	.050	.165
100	3	.108	.313	.659	.034	.142	.427
100	6	.130	.351	.678	.044	.166	.445
100	12	.146	.351	.638	.055	.168	.399
100	24	.141	.295	.508	.054	.128	.278
200	0	.067	.116	.226	.017	.033	.081
200	3	.088	.200	.412	.026	.077	.207
200	6	.101	.222	.421	.034	.091	.212
200	12	.111	.217	.380	.041	.090	.183
200	24	.109	.177	.276	.036	.069	.115

to detect a population shift and thus increasing the population shift bias. Hence, different phenomena may work in opposite directions.

The population shift bias also depends on the level of aggregation. If there are very local population shifts, then it is possible to reduce the bias of the ordinary Knox test by combining areas in which the shifts are in opposite directions from the overall average population growth. Taking this one step further, it is worth pointing out that one way to construct an unbiased Knox test is to aggregate data in such a way that each aggregated area has the same population growth curve. In practice, though, this is hard to accomplish because populations aggregated into the same area must be very close to each other for the test to be meaningful. A better way to obtain an unbiased test is proposed in the following section.

An Unbiased Knox Test

To obtain an unbiased version of the Knox test, it is necessary to know the background population and its temporal trends. Using such data, one can obtain random replications of cases generated under the null hypothesis. These replications can then be used for hypothesis testing using the Monte Carlo procedure. Randomizing in proportion to the population size at each time and place adjusts for the population shifts.

In creating an unbiased Knox test, one must be careful as to how to implement the Monte Carlo method. For example, the Monte Carlo approach suggested by Mantel (1) does not work for this purpose. This is because Mantel proposes to randomize cases using random permutations of spatial and temporal observations conditioned on the set of spatial and set of temporal values, rather than randomizing completely new cases from the background population. The former is the preferred way to do the test when there are no population shifts—when it is not necessary to make distributional approximations—but it does not eliminate the population shift bias.

Neither does it work to simply calculate the Knox test statistic X and, in the normal Monte Carlo fashion, compare its values in the real and randomized datasets. Doing so would give a valid unbiased test, but the value of the test statistic would be high due to purely spatial clustering, purely temporal clustering, or temporal trends. Hence, it would no longer be a test for space-time interaction, but instead a test for global space-time clustering, as discussed by Kulldorff (19).

A way to eliminate the population shift bias and at the same time retain the space-time interaction test is as follows:

1. Generate random datasets for which each random replication has the same number of cases as the real data. The location and time of each case should be random, with probability proportional to the population size for that location and time or to the expected number of cases under the null hypothesis, adjusted for potential confounders such as age.
2. Calculate the test statistic X for the real and random datasets.
3. For each dataset, normalize X using the Barton-David-based approximation:

$$N(X) = \frac{X - E[X|N_t, N_s]}{V[X|N_t, N_s, N_{2s}, N_{2t}]}$$

   This is necessary because N_t, N_s, N_{2s}, and N_{2t} change in each simulated dataset.
4. Rank $N(X)$ for the real and random datasets. If the former is among the 5% highest, reject the null hypothesis of no space-time interaction at the 5% significance level. The corresponding simulated p-value is $R/(REP+1)$, where R is the rank of $N(X)$ from the real dataset and REP is the number of Monte Carlo replications.

For the third step we chose to use the Barton-David-based approximation. Using the Poisson-based approximation will also give an unbiased test. Because only the relative rank is of interest, the accuracy of the approximation is unimportant as long as the ranking it creates is unchanged. The Monte Carlo option for approximating the distribution of X is less practical, as choosing it would mean running one Monte Carlo simulation embedded within another, quite a time-consuming task even for a computer.

Table 5 shows the application of the unbiased Knox test to lung cancer in New Mexico from 1973 to 1991. These data were collected by the New Mexico Tumor Registry for the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program. Table 6 presents the unbiased Knox test as applied to all types of childhood leukemia in Sweden from 1973 to 1994. In both cases, 19,999 Monte Carlo replications were performed. The resulting p-values are given for a range of spatial and temporal critical distances. For comparison, the nominal but biased p-values using the Poisson and Barton-David approximations are given in parentheses.

For the lung cancer data, the unbiased Knox test gives no evidence of any space-time interaction. In contrast, when the population shift bias is not adjusted for, some of the p-values are very small, giving a false impression of space-time interaction. For the leukemia data, out of 30 tests for different critical distances, 8 are significant at the 0.05 level when the unbiased Knox test is used. This may indicate some level of space-time interaction, but it is hard to judge because there is considerable multiple testing involved. This is discussed in the next section.

Table 5 Unbiased p-values When the Knox Test Is Applied to 9,254 Cases of Lung Cancer in New Mexico, 1973–1991, Using Different Critical Distances

	months				
	0	3	6	12	24
0 km	**.361**	**.796**	**.910**	**.888**	**.895**
	(.092/.125)	(.203/.345)	(.540/.512)	(.452/.488)	(.749/.552)
50 km	**.187**	**.396**	**.551**	**.555**	**.696**
	(.005/.012)	(<.0001/.010)	(<.0001/.024)	(<.0001/.030)	(<.0001/.117)
100 km	**.482**	**.423**	**.366**	**.418**	**.544**
	(.016/.027)	(<.0001/.001)	(<.0001/.0004)	(<.0001/.001)	(<.0001/.014)
200 km	**.228**	**.174**	**.138**	**.196**	**.231**
	(.032/.026)	(<.0001/.002)	(<.0001/.0007)	(<.0001/.003)	(<.0001/.012)

Note: In parentheses are the biased p-values from the ordinary Knox test using the Poisson and Barton-David approximations (Poisson/Barton-David). Adjusting for the multiple testing, the unbiased combined p-value is .472.

An Unbiased Combined Knox Test

When the ordinary Knox test is applied, a key feature is the choice of critical distances. Because the scale at which clustering may exist is often unknown, the test has often been applied for a whole range of possible values (e.g., Gilman and Knox, 1995 [14]). This is valuable for estimating the scale of clustering, but it also introduces multiple testing, and if the test is significant for some critical distances but not for others, as in Table 6, then the result is hard to interpret. One solution is to apply some Bonferroni-type adjustment, but because the different tests for different critical distances are statistically dependent, such a procedure is overly conservative and is not commonly used. Using the same basic idea as Baker (13), one can obtain an unbiased combined Knox test as follows.

Table 6 Unbiased p-values When the Knox Test Is Applied to 1,592 Cases of Childhood Leukemia in Sweden, 1973–1994, Using Different Critical Distances

	months				
	0	3	6	12	24
0 km	**.077**	**.135**	**.295**	**.282**	**.074**
	(.061/.061)	(.119/.116)	(.255/.249)	(.246/.236)	(.055/.040)
2 km	**.113**	**.125**	**.238**	**.188**	**.040**
	(.097/.097)	(.109/.106)	(.200/.194)	(.157/.146)	(.028/.018)
5 km	**.224**	**.028**	**.077**	**.423**	**.369**
	(.210/.210)	(.020/.019)	(.058/.054)	(.359/.353)	(.292/.277)
10 km	**.611**	**.049**	**.134**	**.293**	**.398**
	(.609/.609)	(.038/.036)	(.108/.105)	(.250/.247)	(.341/.340)
20 km	**.851**	**.029**	**.028**	**.020**	**.143**
	(.840/.842)	(.020/.020)	(.018/.018)	(.009/.011)	(.100/.121)
50 km	**.362**	**.091**	**.054**	**.033**	**.023**
	(.360/.357)	(.084/.080)	(.042/.041)	(.019/.022)	(.007/.015)

Note: In parentheses are the biased p-values from the ordinary Knox test using the Poisson and Barton-David approximations (Poisson/Barton-David). Adjusting for the multiple testing, the unbiased combined p-value is .237.

1. For the real and random datasets, calculate the test statistic X_d for each of several combinations of critical distances.
2. For each choice of critical distances, calculate the normalized test statistic $N(X_d)$ as described in "An Unbiased Knox Test."
3. For each dataset, select the maximum value of $N(X_d)$ taken over all sets of critical distances, $M=\max_d N(X_d)$.
4. Rank the maximum values M coming from the real and random datasets. If the former is among the 5% highest, reject the null hypothesis of no space-time interaction at the 5% significance level. The corresponding simulated p-value is as before—$R/(REP+1)$, where R is the rank of M from the real dataset and REP is the number of Monte Carlo replications.

For the Swedish childhood leukemia data presented in Table 6, the p-value for the unbiased combined Knox test is 0.237. This indicates that there was no significant space-time interaction of childhood leukemia in Sweden during the period 1973–1994. From an epidemiological viewpoint, though, it is not necessarily the union of all types of leukemia that is of primary interest in a space-time analysis. More detailed analyses by subgroup will be presented in a medicine-oriented paper.

A combined Knox test can be seen not only as a way to account for the multiple testing of several Knox tests, but also as a test in itself to be compared with other space-time interaction tests. Some of these, including Mantel (1), were proposed precisely to avoid the arbitrariness in the choice of critical distances. They are not the same as the combined Knox test, though.

Mantel (1) and Diggle et al. (12) sum up the value of several Knox tests and use the combined sum as an omnibus test statistic. Diggle et al. do the summation for a finite set of critical distances, while Mantel uses a general function leading to continuous summation (integration) if the function is continuous, and to the ordinary Knox test if a dichotomous indicator function is used. The combined Knox test, on the other hand, picks the maximum rather than the sum over a finite set of critical distances.

The choice of method depends on the set of alternative hypotheses for which the user wants to maximize the statistical power. An advantage of the approaches taken by Mantel and Diggle et al. is that they model a gradual decrease in the strength of space-time clustering with increasing distance. A drawback is that the relative strengths at different distances have to be specified a priori. The combined Knox test, on the other hand, models an abrupt cutoff point just like the ordinary Knox test, in which the strength of space-time clustering is constant within the critical distance and zero outside. Unlike the Knox test, though, the critical distances do not need to be specified a priori, and unlike the Mantel and Diggle et al. tests, the relative strengths of clustering at different distances need not be specified. This has two advantages. It is not necessary to limit the scale of space-time interaction to be tested for, and the result provides not only an overall p-value but also, if the result is significant, an indication of the scale at which the space-time interaction operates.

Discussion

In looking at the population shift bias of space-time interaction tests, we have focused on the Knox method because it is the method most widely used for epidemiological data. Such bias is also present in other space-time interaction tests, proposed by David and Barton (10), Mantel (1), Pike and Smith (11), Diggle et al. (12), Jacquez (4), and Baker (13). The Mantel test, and even more so the Jacquez test, have been shown to have higher power than the Knox test for certain alternative hypotheses (4). Ironically, this also means that the population shift bias is higher, because a test's population shift bias is simply its power to detect the space-time interaction inherent in the population distribution. Fortunately, the procedure for constructing the unbiased Knox test can also be used for the Mantel and Jacquez tests, in the same simple fashion.

An unbiased combined Jacquez test would be especially attractive. Rather than using fixed geographic distances as Knox (5), Mantel (1), and Diggle et al. (12) have done, Jacquez (4) defines distances in terms of nearest neighbors, so that cases 1 kilometer apart are considered to be close to each other in a rural area but not necessarily so in a densely populated city. This increases the power when there is space-time interaction in less-populated areas.

No matter which space-time interaction test is used, it would have been ideal to show that, for practical purposes, the population shift bias is more or less irrelevant. Unfortunately, that is not the case (see "Estimation of the Population Shift Bias"). This leads to two questions: How do we do this type of analysis in the future? How do we interpret past results in light of the bias that may be associated with them?

To perform an unbiased test for space-time interaction in an area, we need underlying population data for that area. These data are sometimes harder to get than the case data. If a proper test is to be performed, there is no way around this, but in some cases there is a shortcut. The ordinary space-time interaction tests are all liberal. Therefore,

we know that if there is no significant space-time interaction using the ordinary test, then space-time interaction will not be significant according to the unbiased version. This suggests a two-stage procedure. First, collect only the case data and use one of the ordinary space-time interaction tests. If the result is non-significant, then there is no need to obtain the population data and the negative results can be published as such. If the result is significant, though, the population shift bias may be affecting it. It is then important to obtain population data and apply the unbiased version before making any conclusions.

Caution should be used in interpreting results that have already been published. If a result is non-significant, then it is fine. If the study period was only one or two years, the population shift bias is probably not a major problem because differential changes in population sizes did not have much chance to accumulate. For datasets spanning 10 or 20 years, though, there is really no way of knowing how reliable the results are without reanalyzing the data using an unbiased approach. For any past results that are considered important from an etiological or public health standpoint, we recommend that the data be reanalyzed using the unbiased version of any of the space-time interaction tests.

Acknowledgments

Valuable suggestions from Geoffrey Jacquez are gratefully acknowledged.

References

1. Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27:209–20.
2. Williams GW. 1984. Time-space clustering of disease. In: *Statistical methods for cancer studies.* Ed. RG Cornell. New York: Marcel Dekker.
3. Chen R, Mantel N, Klingberg MA. 1984. A study of three techniques for time-space clustering in Hodgkin's disease. *Statistics in Medicine* 3:173–84.
4. Jacquez GM. 1996. A *k* nearest neighbor test for space-time interaction. *Statistics in Medicine* 15:1935–49.
5. Knox G. 1964. The detection of space-time interactions. *Applied Statistics* 13:25–9.
6. Wartenberg D, Greenberg M. 1994. Personal communication.
7. Alexander FE. 1992. Space-time clustering of childhood acute lymphoblastic leukemia: Indirect evidence for a transmissible agent. *British Journal of Cancer* 65:589–92.
8. Petridou E, Revinthi K, Alexander FE, Haidas S, Koliouskas D, Kosmidis H, Piperopoulou F, Tzortzatou F, Trichopoulos D. 1996. Space-time clustering of childhood leukemia in Greece: Evidence supporting a viral etiology. *British Journal of Cancer* 73:1278–83.
9. Klauber MR, Mustacchi P. 1970. Space-time clustering of childhood leukemia in San Francisco. *Cancer Research* 30:1969–73.
10. David FN, Barton DE. 1966. Two space-time interaction tests for epidemicity. *British Journal of Preventive Social Medicine* 20:44–8.
11. Pike MC, Smith PG. 1968. Disease clustering: A generalization of Knox's approach to the detection of space-time interactions. *Biometrics* 24:541–56.

12. Diggle P, Chetwynd AG, Häggkvist R, Morris SE. 1995. Second-order analysis of space-time clustering. *Statistical Methods in Medical Research* 4:124–36.
13. Baker RD. 1996. Testing for space-time clusters of unknown size. *Journal of Applied Statistics* 23:543–54.
14. Gilman EA, Knox EG. 1995. Childhood cancers: Space-time distribution in Britain. *Journal of Epidemiology and Community Health* 49:158–63.
15. Barton DE, David FN. 1966. The random intersection of two graphs. In: *Research papers in statistics: Festschrift for Jerzy Neyman.* Ed. FN David. London: John Wiley & Sons. 445–59.
16. Aldrich TE, Drane JW. 1993. *Cluster, v. 3.0: A program for identifying and analyzing the spatial and temporal structure of chronic disease patterns.* Atlanta, GA: Agency for Toxic Substances and Disease Registry.
17. Dwass M. 1957. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* 28:181–7.
18. Jacquez GM. 1994. *Stat!: Statistical software for the clustering of health events.* Ann Arbor, MI: BioMedware.
19. Kulldorff M. 1998. Statistical methods for spatial epidemiology: Tests for randomness. In: *GIS and health in Europe.* Ed. A Gatrell, M Löytönen. London: Taylor & Francis.